Statistica Sinica Preprint No: SS-2023-0201	
Title	An Empirical Bayes Regression for Multi-tissue Gene Expression Prediction
Manuscript ID	SS-2023-0201
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0201
Complete List of Authors	Fei Xue and Hongzhe Li
Corresponding Authors	Hongzhe Li
E-mails	hongzhe@upenn.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

AN EMPIRICAL BAYES REGRESSION FOR MULTI-TISSUE GENE EXPRESSION PREDICTION

Fei Xue[1] and Hongzhe Li[2]

[1] *Purdue University and* [2] *University of Pennsylvania*

Abstract: The Genotype-Tissue Expression (GTEx) project collects samples from multiple human tissues to study the relationship between genetic variation or single nucleotide polymorphisms (SNPs) and gene expression in each tissue. However, most existing eQTL analyses only focus on single tissue information. In this paper, we develop a multi-tissue method that improves prediction of gene expression based on cis-SNPs by borrowing information across tissues. Specifically, we propose an empirical Bayes regression model for SNP-expression association using data from multiple tissues. To allow the effects of SNPs to vary greatly among tissues, we use a mixture distribution as the prior, which is a mixture of a multivariate Gaussian distribution and a Dirac mass at zero. We show that the proposed estimator of the cis-SNP effects on gene expression asymptotically achieves the minimum Bayes risk among all estimators. Analyses of the GTEx data show that our proposed method is superior to existing methods in terms of prediction accuracy for gene expression using cis-SNPs in testing sets.

Key words and phrases: Bayes risk, Data integration, Missing data, Mixture model

1. Introduction

Genome-wide association studies (GWAS) have successfully associated single nucleotide polymorphisms (SNPs) with complex human traits (Uffelmann et al., 2021). However, there are still problems in statistical power and interpretation of GWAS results due to complexity

of linkage disequilibrium (LD) and gene regulation (Boyle et al., 2017). To alleviate these problems, a popular approach is the transcriptome-wide association study (TWAS) that integrates the SNP-trait association with SNP-based prediction of gene expression (Wainberg et al., 2019). Specifically, TWAS first predicts expression levels using SNPs, and then tests whether the predicted values are associated with human traits. In this paper, we focus on the first part in TWAS and aim to improve the SNP-based prediction of gene expression.

Many large data sets have been generated for such genetics of gene expression studies for various tissues, which have provided important insights into gene regulations. Among these studies, the Genotype-Tissue Expression (GTEx) project aims to characterize variation in gene expression levels across individuals and diverse tissues, many of which are not easily accessible (Consortium et al., 2017). The project found that local genetic variation affects gene expression levels for the majority of genes, and identified inter-chromosomal genetic effects for a small number of genes and loci. Such expression quantitative trait loci (eQTL) analyses provide important insights into genetic regulation of gene expressions. The GTEx data sets have also been applied to impute gene expression levels based on genetic variants data and the imputed gene expressions are subsequently used in TWAS analysis (Gamazon et al., 2015; Gusev et al., 2016; Hu et al., 2019).

However, small sample sizes of many studies, e.g., only hundreds of samples for each tissue type in the GTEx study, often limits the prediction accuracy of gene expression levels based on genotype data. To date, most eQTL studies have considered the association between genetic variation and expression in a single tissue (Brem et al., 2005; Stranger et al., 2007; Stegle et al., 2012). Multi-tissue eQTL analysis has the potential to improve the findings of single tissue analyses by borrowing strength across tissues and to elucidate

the genetic basis of the difference in expressions between tissues (Flutre et al., 2013; Sul et al., 2013; Duong et al., 2017; Li et al., 2018). However, these methods only focus on testing the association between gene expression and a single SNP.

More recently, several multi-tissue multi-locus models have been proposed in Morgante et al. (2023); Shi et al. (2020); Hu et al. (2019); Molstad et al. (2021). For example, Hu et al. (2019) and Molstad et al. (2021) considered fixed effects and integrated information across multiple tissues using a group-lasso penalty on effects of each SNP in all the tissues. Shi et al. (2020) adopted a factorizable assumption to borrow information across tissues. Specifically, they assumed that the effects of SNPs on the gene expression in multiple tissues can be decomposed into genomic variant-dependent and tissue-dependent components, where the genomic variant-dependent one contains shared information across tissues. Nevertheless, these methods do not account for potential large variation of genetic effects across tissues, i.e., where SNPs are associated with expression in only a subset of tissues for some genes. In such circumstances, estimators of shared effects from all the tissues will be biased by inclusion of tissues where SNPs are not involved with expression. In fact, gene expression and regulation are often tissue-specific (Sonawane et al., 2017), and some genes express uniquely in certain tissues (Dezső et al., 2008), which is important to discovery of new drugs and biomarkers of tissue-targeted diseases.

There are also several empirical Bayes methods developed for multi-tissue analysis. For instance, for each SNP, Kim et al. (2022) assumed a finite mixture of univariate normal distributions as a prior of the effect in each tissue, while Morgante et al. (2023) adopted a mixture of multivariate normal distributions as a prior of effects in multiple tissues and developed Mr.Mash to captures similarity of effects across tissues through the covariance

matrices in the multivariate normal distributions. Although the mixture of multivariate normals prior is flexible, Mr.Mash assumes that the covariance matrices are pre-specified, and treats the mixture weights as parameters to be estimated from the data. Lastly, Wang and Zhao (2021) proposed a nonparametric empirical Bayes method, which assumes a unknown prior distribution of effects and estimate it nonparametrically from data.

In this paper, we develop an alternative empirical Bayes regression model for SNP-based gene-expression prediction and SNP-expression association analysis using data from multiple tissues, allowing for potential large variation across tissues and joint effects of multiple genetic variants on gene expressions. Our model serves two purposes. One is to predict gene expression using corresponding cis-SNPs for each gene and tissue, where cis-SNPs of a given gene are SNPs located either anywhere within the region from 1 Mb upstream and downstream the gene. The other is to test whether these cis-SNPs are associated with the gene expression (see Section B of the supplementary materials), that is, whether the gene is an eGene whose expression level is related to at least one cis-SNP (Duong et al., 2016).

To achieve these goals, for each gene, we construct tissue-specific linear regression models with expression level of the gene as the response and its corresponding cis-SNPs as predictors. To borrow information across different tissues, we propose an empirical Bayes estimator for the regression coefficients based on a mixture prior distribution. We adopt the posterior mean of the coefficients for estimation, and estimate the prior parameters in the the posterior mean through maximizing the marginal likelihood of the gene expression values in all tissues based on the expectation–maximization (EM) algorithm (Dempster et al., 1977). We then extract evidence of whether the cis-SNPs are relevant to the gene expression from the data by calculating posterior probabilities and Bayes factor of the hypotheses.

The main contributions of the proposed method are as follows. First, we extract shared information across tissues via the common prior distribution of the regression coefficients in tissue-specific models. We propose to combine the information in single tissue and the shared information in the prior distribution using the empirical Bayes estimator. In theory, we show that the proposed estimator is superior to the traditional ordinary least squared (OLS) estimator for a given tissue in terms of the Bayes risk and the mean squared error.

Second, we incorporate situations where all the cis-SNPs are irrelevant to the gene expression in some tissues through a mixture prior distribution of coefficients. One of the two components is exactly a zero vector, while the other one has non-zero mean representing shared information across tissues with non-zero effects. In this way, we can test whether a gene is an eGene in a specific tissue based on the posterior probabilities of the assignments for the two components for a given tissue. Through our analysis of the GTEx data in Section 7, we show that the proposed method outperforms existing methods in terms of prediction of gene expression. Moreover, although genetic effects on expression are extensively shared among some tissues, we found that effect sizes can still vary greatly across tissues, illustrated in Figure E.1 in the supplementary materials.

2. An empirical Bayes regression model for SNP-expression association across multiple tissues

2.1 Empirical Bayes regression

In this section, we link the SNP genotypes with gene expression in each tissue by a tissue-specific linear regression model. Specifically, we let $\boldsymbol{Y}$ denote a $n \times m$ matrix consisting of expression values of a gene in m tissues of n samples and $\boldsymbol{X}$ denote a $n \times p$ constant

matrix consisting of cis-SNPs for the gene, where n is the number of total individuals, m is the number of tissues in the data, and p is the number of cis-SNPs. For the t-th tissue, the tissue-specific linear regression model is

$$\boldsymbol{Y}^{(t)} = \boldsymbol{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{\varepsilon}^{(t)}, \tag{2.1}$$

where $\boldsymbol{Y}^{(t)}$ denotes the t-th column in $\boldsymbol{Y}$, $\boldsymbol{\beta}^{(t)}$ is a p-dimensional coefficient vector for the t-th tissue, and $\boldsymbol{\varepsilon}^{(t)} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ is the error term with parameter $\sigma > 0$ and independent of $\boldsymbol{X}\boldsymbol{\beta}^{(t)}$. We also assume that $\boldsymbol{\varepsilon}^{(t)}$ for $t = 1, \ldots, m$ are independent. The ordinary least squares (OLS) estimator of $\boldsymbol{\beta}^{(t)}$ based on information in a single tissue is

$$\hat{\boldsymbol{\beta}}^{(t)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^{(t)} = \boldsymbol{\beta}^{(t)} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}^{(t)}, \tag{2.2}$$

Then $\hat{\boldsymbol{\beta}}^{(t)} \mid \boldsymbol{\beta}^{(t)} \sim N_p(\boldsymbol{\beta}^{(t)}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

To borrow information across tissues, we assume that coefficient vectors $\boldsymbol{\beta}^{(t)}$ over all the tissues are random and have a common prior distribution. This common prior contains shared effects of cis-SNPs across tissues. However, the effects of cis-SNPs in different tissues could vary greatly. Especially, the cis-SNPs could be "inactive" and have no effects on gene expression in some tissues, which can not contribute to the shared effects.

To accommodate this possibility, we define a random indicator $I^{(t)}$, which follows a Bernoulli distribution with probability $\tau_1 \in (0, 1)$, to reflect the status of $\boldsymbol{\beta}^{(t)}$. We assign a mixture prior distribution with two mixture components for $\boldsymbol{\beta}^{(t)}$, that is,

$$\boldsymbol{\beta}^{(t)} \mid I^{(t)} = 1 \sim N_p(\boldsymbol{\beta}, \eta(\boldsymbol{X}^T\boldsymbol{X})^{-1}), \tag{2.3}$$

$$\boldsymbol{\beta}^{(t)} \mid I^{(t)} = 0 \equiv \boldsymbol{0}, \tag{2.4}$$

independently for $t = 1, \ldots, m$, where $\eta > 0$ is a parameter. Here the prior mean $\boldsymbol{\beta}$ represents shared effects across tissues, and $I^{(t)}$ is a latent configuration variable reflecting

the status of $\boldsymbol{\beta}^{(t)}$. When $I^{(t)} = 0$, the cis-SNPs are "inactive" and have no effects on the gene expression $\boldsymbol{Y}^{(t)}$. In contrast, when $I^{(t)} = 1$, the cis-SNPs are "active" and the effects $\boldsymbol{\beta}^{(t)}$ follows a multivariate normal distribution with mean $\boldsymbol{\beta}$. We refer to our proposed method as "multivariate Empirical Bayes method with mixture Prior" or "mEBmix".

We provide the posterior probabilities of $I^{(t)}$ and the posterior mean of $\boldsymbol{\beta}^{(t)}$ in the following proposition. Let $\psi(\boldsymbol{z}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ denote the density function of the multivariate normal distribution $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

Proposition 1. The posterior means of $\boldsymbol{\beta}^{(t)}$ given $I^{(t)}$ are

$$E(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{Y}^{(t)}, I^{(t)} = 1) = \left(\frac{1}{\eta} + \frac{1}{\sigma^2}\right)^{-1} \left(\frac{\boldsymbol{\beta}}{\eta} + \frac{\hat{\boldsymbol{\beta}}^{(t)}}{\sigma^2}\right),$$

and $E(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{Y}^{(t)}, I^{(t)} = 0) = 0$. The posterior probabilities of $I^{(t)}$ are

$$P(I^{(t)} = 1 \mid \boldsymbol{Y}) = h_1(\boldsymbol{Y}^{(t)}; \tau_1, \boldsymbol{\beta}, \eta, \sigma^2),$$

and $P(I^{(t)} = 0 \mid \boldsymbol{Y}) = 1 - h_1(\boldsymbol{Y}^{(t)}; \tau_1, \boldsymbol{\beta}, \eta, \sigma^2)$, where

$$h_1(\boldsymbol{Y}^{(t)}; \tau_1, \boldsymbol{\beta}, \eta, \sigma^2) = \frac{\tau_1 \psi(\boldsymbol{Y}^{(t)}; \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n + \eta \boldsymbol{H})}{\tau_1 \psi(\boldsymbol{Y}^{(t)}; \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n + \eta \boldsymbol{H}) + \tau_0 \psi(\boldsymbol{Y}^{(t)}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)},$$

with $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\tau_0 = 1 - \tau_1$. Thus, the posterior mean of $\boldsymbol{\beta}^{(t)}$ is

$$E(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{Y}) = E(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{Y}^{(t)}) = h_1(\boldsymbol{Y}^{(t)}; \tau_1, \boldsymbol{\beta}, \eta, \sigma^2) \left(\frac{1}{\eta} + \frac{1}{\sigma^2}\right)^{-1} \left(\frac{\boldsymbol{\beta}}{\eta} + \frac{\hat{\boldsymbol{\beta}}^{(t)}}{\sigma^2}\right). \quad (2.5)$$

According to the Proposition 1, the posterior mean of $\boldsymbol{\beta}^{(t)}$ is a weighted average of the OLS estimator in Equation (2.2) and the mean of the prior distribution in Equation (2.3), which combines the information in the t-th tissue and the shared information across tissues in the prior. The first equality in (2.5) follows from the fact that $\boldsymbol{\beta}^{(t)}$ and columns in $\boldsymbol{Y}$ other than $\boldsymbol{Y}^{(t)}$ are conditionally independent given $\boldsymbol{Y}^{(t)}$. The weights in (2.5) are related to the variance of the error term and the variance of the prior distribution.

Moreover, we calculate Bayes factors (BF) and posterior odds ratios in Section B of the supplementary materials to determine whether the cis-SNPs is relevant to the gene expression or not. In Section 3, we estimate the unknown parameters in Equation (2.5) via an expectation-maximization (EM) algorithm (Dempster et al., 1977).

3. Parameter estimation and EM algorithm

In this section, we provide a detailed iterative algorithm to estimate the parameters in the mixture prior distribution through maximizing the likelihood of the data. Specifically, we exploit an EM algorithm (Dempster et al., 1977) to find the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ consists of all the parameters in the model, that is, $\boldsymbol{\theta} = (\tau_1, \tau_0, \boldsymbol{\beta}, \eta, \sigma^2)$. Each iteration consists of an expectation step and a maximization step. Suppose that we have both $\boldsymbol{Y}^{(t)}$ and $\boldsymbol{I}^{(t)}$ for each $t = 1, \ldots, m$. We refer to $\{\boldsymbol{Y}, I^{(1)}, \ldots, I^{(m)}\}$ as the complete data. The complete-data likelihood is

$$p(\boldsymbol{Y}, I^{(1)}, \ldots, I^{(m)}; \tau_1, \tau_0, \boldsymbol{\beta}, \eta, \sigma^2) = \prod_{t=1}^{m} \prod_{s=0}^{1} \left\{ \tau_s g_s(\boldsymbol{Y}^{(t)}; \boldsymbol{\beta}, \eta, \sigma^2) \right\}^{\mathbb{I}(I^{(t)}=s)},$$

where $\mathbb{I}(\cdot)$ is an indicator function, $g_0(\boldsymbol{Y}^{(t)}; \boldsymbol{\beta}, \eta, \sigma^2) = g_0(\boldsymbol{Y}^{(t)}; \sigma^2) = \psi(\boldsymbol{Y}^{(t)}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ and $g_1(\boldsymbol{Y}^{(t)}; \boldsymbol{\beta}, \eta, \sigma^2) = \psi(\boldsymbol{Y}^{(t)}; \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n + \eta \boldsymbol{H})$ denote the likelihoods of $\boldsymbol{Y}^{(t)}$ when $I^{(t)} = 0$ and $I^{(t)} = 1$, respectively.

In the expectation step, since we typically do not observe $\{I^{(1)}, \ldots, I^{(m)}\}$ in practice, given the current estimate $\boldsymbol{\theta}_{(k)}$ at the k-th iteration, we first calculate the posterior distribution of $I^{(t)}$

$$T_{s,(k)}^{(t)} = P(I^{(t)} = s \mid \boldsymbol{Y}, \boldsymbol{\theta}_{(k)}) = \frac{\tau_{s,(k)} g_s(\boldsymbol{Y}^{(t)}; \boldsymbol{\beta}_{(k)}, \eta_{(k)}, \sigma_{(k)})}{\tau_{1,(k)} g_1(\boldsymbol{Y}^{(t)}; \boldsymbol{\beta}_{(k)}, \eta_{(k)}, \sigma_{(k)}) + \tau_{0,(k)} g_0(\boldsymbol{Y}^{(t)}; \sigma_{(k)})}$$

for $s = 0, 1$, where $\tau_{s,(k)}$, $\boldsymbol{\beta}_{(k)}$, $\eta_{(k)}$, and $\sigma_{(k)}$ denote estimates of τ, $\boldsymbol{\beta}$, η, and σ at the k-th

iteration. Moreover, we calculate the expectation of the complete-data log-likelihood under the posterior distribution of the latent variables $\{I^{(1)}, \ldots, I^{(m)}\}$:

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{(k)}) &= E_{I^{(1)},\ldots,I^{(m)}|\boldsymbol{Y},\boldsymbol{\theta}_{(k)}} \left[\log p(\boldsymbol{Y}, I^{(1)}, \ldots, I^{(m)}; \boldsymbol{\theta})\right] \\ &= \sum_{t=1}^{m} \sum_{s=0}^{1} T_{s,(k)}^{(t)} \left\{\log \tau_s + \log g_s(\boldsymbol{Y}^{(t)}; \boldsymbol{\beta}, \eta, \sigma^2)\right\}. \end{aligned}$$

In the maximization step, we maximize this expectation to determine the next estimate for all the parameters. The maximizer of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{(k)})$ consists of

$$\tau_{s,(k+1)} = \frac{\sum_{t=1}^{m} T_{s,(k)}^{(t)}}{\sum_{t=1}^{m} \left\{T_{0,(k)}^{(t)} + T_{1,(k)}^{(t)}\right\}} \quad \text{for } s = 1, 0, \quad \boldsymbol{\beta}_{(k+1)} = \frac{\sum_{t=1}^{m} T_{1,(k)}^{(t)} \hat{\boldsymbol{\beta}}^{(t)}}{\sum_{t=1}^{m} T_{1,(k)}^{(t)}},$$

$$\sigma_{(k+1)}^2 = \frac{\sum_{t=1}^{m} \left(\boldsymbol{Y}^{(t)}\right)^T \left(\boldsymbol{Y}^{(t)}\right) - \sum_{t=1}^{m} T_{1,(k)}^{(t)} \left(\boldsymbol{Y}^{(t)}\right)^T \boldsymbol{H} \left(\boldsymbol{Y}^{(t)}\right)}{mn - p \sum_{t=1}^{m} T_{1,(k)}^{(t)}},$$

and

$$\eta_{(k+1)} = \frac{\sum_{t=1}^{m} T_{1,(k)}^{(t)} \left(\boldsymbol{Y}^{(t)} - \boldsymbol{X}\boldsymbol{\beta}_{(k+1)}\right)^T \boldsymbol{H} \left(\boldsymbol{Y}^{(t)} - \boldsymbol{X}\boldsymbol{\beta}_{(k+1)}\right)}{p \sum_{t=1}^{m} T_{1,(k)}^{(t)}} - \sigma_{(k+1)}^2.$$

In this way, we derive closed-form expression updates for each iteration, which are straightforward to compute. In addition, this EM algorithm converges since each iteration does increase the likelihood of observed data.

4. Statistical properties

In this section, we provide the asymptotic results of the proposed estimator in terms of a Bayes risk function. Specifically, we define the Bayes risk function of an estimator $\boldsymbol{\delta}_m(\boldsymbol{Y}) \in \mathcal{D}_m$ for $\boldsymbol{\beta}^{(t)}$ as

$$R_m(\boldsymbol{\delta}_m) = \int l\left(\boldsymbol{\beta}^{(t)}, \boldsymbol{\delta}_m(\boldsymbol{Y})\right) \prod_{i=1}^{m} \left\{p(\boldsymbol{Y}^{(i)} \mid \boldsymbol{\beta}^{(i)}) p(\boldsymbol{\beta}^{(i)}) d\boldsymbol{Y}^{(i)} d\boldsymbol{\beta}^{(i)}\right\},$$

where $\mathcal{D}_m$ is the set consisting of all available estimators of $\boldsymbol{\beta}^{(t)}$, and

$$l\left(\boldsymbol{\beta}^{(t)}, \widetilde{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y})\right) = \left\{\widetilde{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y}) - \boldsymbol{\beta}^{(t)}\right\}^T \Delta \left\{\widetilde{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y}) - \boldsymbol{\beta}^{(t)}\right\},$$

is a squared error loss function with a positive definite matrix Δ. Let $\widetilde{\tau}_0$, $\widetilde{\tau}_1$, $\widetilde{\eta}$, $\widetilde{\sigma}$, and $\widetilde{\boldsymbol{\beta}}$ be the MLEs of τ_0, τ_1, η, σ, and $\boldsymbol{\beta}$, respectively. Then, by Proposition 1, the proposed empirical Bayes estimator of $\boldsymbol{\beta}^{(t)}$ for the t-th tissue is

$$\widetilde{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y}) \;=\; \frac{\widetilde{\eta}\widetilde{\sigma}^2 \cdot h_1(\boldsymbol{Y}^{(t)}; \widetilde{\tau}_1, \widetilde{\boldsymbol{\beta}}, \widetilde{\eta}, \widetilde{\sigma})}{\widetilde{\eta} + \widetilde{\sigma}^2} \left(\frac{\widetilde{\boldsymbol{\beta}}}{\widetilde{\eta}} + \frac{\hat{\boldsymbol{\beta}}^{(t)}}{\widetilde{\sigma}^2}\right).$$

If τ_1, η, σ, and $\boldsymbol{\beta}$ are known, then we can use

$$\bar{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y}) = E(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{Y}^{(t)}) = \frac{\eta\sigma^2 \cdot h_1(\boldsymbol{Y}^{(t)}; \tau_1, \boldsymbol{\beta}, \eta, \sigma^2)}{\eta + \sigma^2} \left(\frac{\boldsymbol{\beta}}{\eta} + \frac{\hat{\boldsymbol{\beta}}^{(t)}}{\sigma^2}\right) \tag{4.6}$$

as an estimator for $\boldsymbol{\beta}^{(t)}$. We refer to the $\bar{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y})$ as an oracle estimator. Let $\varphi(\alpha) = \int_{\boldsymbol{x} \notin \mathcal{B}(\mathbf{0},\alpha)} \psi(\boldsymbol{x}; \mathbf{0}, \boldsymbol{I}_p) d\boldsymbol{x}$, where $\mathcal{B}(\mathbf{0}, \alpha)$ represents the ball centered at $\mathbf{0}$ with radius α. We provide the theoretical results for the oracle estimator in the following theorem.

Theorem 1. If τ_1, η, σ, and $\boldsymbol{\beta}$ are known, then the oracle estimator $\bar{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y})$ in Equation (4.6) is optimal, that is,

$$R_m(\bar{\boldsymbol{\beta}}^{(t)}) = \inf_{\boldsymbol{\delta}_m \in \mathcal{D}_m^*} R_m(\boldsymbol{\delta}_m)$$

for each $1 \le t \le m$, where $\mathcal{D}_m^*$ is the set consisting of all available estimators of $\boldsymbol{\beta}^{(t)}$ with known τ_1, η, σ, and $\boldsymbol{\beta}$. In addition, for each $1 \le t \le m$,

$$\begin{aligned} R_m(\hat{\boldsymbol{\beta}}^{(t)}) - R_m(\bar{\boldsymbol{\beta}}^{(t)}) \;\ge\; & \frac{\sigma^2\alpha^2\lambda_{\min}(\Delta)}{\eta + \sigma^2} \left(\tau_1\varphi\left[\lambda_{\max}\{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}\}(2\|\boldsymbol{\beta}\|_2 + \alpha)/(\sigma^2 + \eta)^{1/2}\right]\right. \\ & \left. + \tau_0\varphi\left[\lambda_{\max}\{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}\}(\|\boldsymbol{\beta}\|_2 + \alpha)/\sigma\right]\right), \end{aligned} \tag{4.7}$$

where α is any positive constant, and $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the largest and smallest eigenvalues, respectively.

Theorem 1 states that the oracle estimator $\bar{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y})$ can achieve the minimum Bayes risk among all estimators based on known prior parameters. The equation (4.7) implies that the oracle estimator is strictly better than the OLS estimator in terms of the Bayes risk function. In the following theorem, we show the convergence of the proposed estimator to the oracle estimator as the total number of tissues goes to infinity.

Theorem 2. The proposed estimator $\widetilde{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y})$ converges in probability to the oracle estimator, that is,

$$\left\| \widetilde{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y}) - \bar{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y}) \right\|_2 \xrightarrow{p} 0,$$

as $m \to \infty$.

As shown in Theorem 2, when we have more tissues, the proposed estimator gets closer to the optimal oracle estimator $\bar{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y})$. In contrast, the OLS estimator $\hat{\boldsymbol{\beta}}^{(t)}$ stays apart from $\bar{\boldsymbol{\beta}}^{(t)}(\boldsymbol{Y})$. This is due to that the proposed estimator borrows cross-tissue information in the estimation of the common prior parameters, while the OLS estimator only uses information in one single tissue. Note that we do not require that the sample size per tissue n goes to infinity in Theorem 2, but we still need $n > p$ since the OLS estimator is involved in the proposed estimator $\widetilde{\boldsymbol{\beta}}^{(t)}$.

5. An empirical Bayes regression model with missing data

In this section, we consider situations where there are missing values in gene expression matrix $\boldsymbol{Y}$, which is motivated by missing tissue samples in the GTEx data. Specifically, for each $1 \leq t \leq m$, let $\boldsymbol{W}^{(t)}$ be a $n \times n$ diagonal matrix with binary diagonal elements $w_{ii}^{(t)}$ for $1 \leq i \leq n$, $w_{ii}^{(t)} = 1$ if and only if $y_i^{(t)}$ is observed, where $y_i^{(t)}$ is the i-th element in $\boldsymbol{Y}^{(t)}$. We assume that each $w_{ii}^{(t)}$ is independent of each other and missing is at random. Then, the

OLS estimator for $\boldsymbol{\beta}^{(t)}$ with missing data is

$$\hat{\boldsymbol{\beta}}_{obs}^{(t)} = (\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{Y}^{(t)} = \boldsymbol{\beta}^{(t)} + (\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{\varepsilon}^{(t)}. \tag{5.8}$$

Compared with the OLS estimator in (2.2), $\hat{\boldsymbol{\beta}}_{obs}^{(t)}$ is constructed only based on subjects whose gene expression levels in the t-th tissue type are observed. We also have $\hat{\boldsymbol{\beta}}_{obs}^{(t)} \mid \boldsymbol{\beta}^{(t)}, \boldsymbol{W}^{(t)} \sim N_p(\boldsymbol{\beta}^{(t)}, \sigma^2(\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1})$.

Following a similar derivation, we can derive the posterior mean of $\boldsymbol{\beta}^{(t)}$ as

$$\begin{aligned} E(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{W}) &= E(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{Y}_{\text{obs}}^{(t)}, \boldsymbol{W}^{(t)}) \\ &= h_2(\boldsymbol{Y}_{\text{obs}}^{(t)}, \boldsymbol{W}^{(t)}; \tau_1, \boldsymbol{\beta}, \eta, \sigma^2)(\boldsymbol{X}^T\boldsymbol{X}/\eta + \boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X}/\sigma^2)^{-1} \\ &\quad \cdot(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}/\eta + \boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{Y}^{(t)}/\sigma^2). \end{aligned} \tag{5.9}$$

where

$$h_2(\boldsymbol{Y}_{\text{obs}}^{(t)}, \boldsymbol{W}^{(t)}; \tau_1, \boldsymbol{\beta}, \eta, \sigma^2) = \frac{\tau_1\psi(\boldsymbol{Y}_{\text{obs}}^{(t)}; \boldsymbol{X}_t\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{n_t} + \eta\boldsymbol{H}_t)}{\tau_1\psi(\boldsymbol{Y}_{\text{obs}}^{(t)}; \boldsymbol{X}_t\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{n_t} + \eta\boldsymbol{H}_t) + \tau_0\psi(\boldsymbol{Y}_{\text{obs}}^{(t)}; \boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_t})}.$$

Compared with the posterior mean in (2.5), the posterior mean of $\boldsymbol{\beta}^{(t)}$ in (5.9) is more complicated since $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X}$ are not exactly the same due to missing values. However, $E(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{W})$ is still a weighted combination of the shared information in $\boldsymbol{\beta}$ and the observed information for the t-th tissue. We adopt the posterior expectation with MLEs of $\boldsymbol{\beta}, \tau_1, \eta, \sigma^2$ as our proposed estimator. To find the MLEs, we provide an EM algorithm in Supplementary Materials A, where the estimation for $\boldsymbol{\beta}$ is calculated based on observed samples in all the tissues. Under the setting with missing data, a Bayes factor and posterior odds ratio can be derived (see Supplemental Materials).

For each tissue t, let $\widetilde{\boldsymbol{\beta}}_{obs}^{(t)}(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{W})$ be the posterior mean given in (5.9) with the parameters estimated using the MLEs, $\widetilde{\tau}_1$, $\widetilde{\eta}$, $\widetilde{\sigma}$, $\widetilde{\boldsymbol{\beta}}$. We define the Bayes risk function of any

estimator $\boldsymbol{\delta}_m(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{W}) \in \bar{\mathcal{D}}_m$ for $\boldsymbol{\beta}^{(t)}$ as

$$R_m^{(obs)}(\boldsymbol{\delta}_m) = \int l\left(\boldsymbol{\beta}^{(t)}, \boldsymbol{\delta}_m(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{W})\right) \prod_{i=1}^{m} \left\{ p(\boldsymbol{W}^{(i)}, \boldsymbol{Y}_{\text{obs}}^{(i)} \mid \boldsymbol{\beta}^{(i)}) p(\boldsymbol{\beta}^{(i)}) d\boldsymbol{Y}_{\text{obs}}^{(i)} d\boldsymbol{W}^{(i)} d\boldsymbol{\beta}^{(i)} \right\},$$

where $\bar{\mathcal{D}}_m$ is the set consisting of all available estimators of $\boldsymbol{\beta}^{(t)}$ with missing data. When η, σ, and $\boldsymbol{\beta}$ are known, let $\bar{\boldsymbol{\beta}}_{obs}^{(t)}(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{W}) = E(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{W})$ be the oracle estimator. Similarly as in Theorem 1, we demonstrate that the proposed estimator $\widetilde{\boldsymbol{\beta}}_{obs}^{(t)}(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{W})$ is strictly better than the OLS estimator $\hat{\boldsymbol{\beta}}_{obs}^{(t)}$ in equation (5.8) in terms of the Bayes risk function via the above oracle estimator in the following theorems.

Theorem 3. If η, σ, and $\boldsymbol{\beta}$ are known, the oracle estimator $\bar{\boldsymbol{\beta}}_{obs}^{(t)}(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{W})$ is optimal, that is,

$$R_m^{(obs)}(\bar{\boldsymbol{\beta}}_{obs}^{(t)}) = \inf_{\boldsymbol{\delta}_m \in \bar{\mathcal{D}}_m^*} R_m^{(obs)}(\boldsymbol{\delta}_m)$$

for each $1 \le t \le m$, where $\bar{\mathcal{D}}_m^*$ is the set consisting of all available estimators of $\boldsymbol{\beta}^{(t)}$ with known η, σ, and $\boldsymbol{\beta}$. In addition, for each $1 \le t \le m$,

$$\begin{aligned} & R_m^{(obs)}(\hat{\boldsymbol{\beta}}_{obs}^{(t)}) - R_m^{(obs)}(\bar{\boldsymbol{\beta}}_{obs}^{(t)}) \qquad (5.10) \\ \ge\ & \kappa \frac{\sigma^2 \alpha^2 \lambda_{\min}(\Delta)}{\eta + \sigma^2} \left(\tau_1 \varphi \left[\lambda_{\max}\{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}\} ((1+\kappa)\|\boldsymbol{\beta}\|_2 + \alpha)/(\sigma^2+\eta)^{1/2} \right] \right. \\ & \left. + \tau_0 \varphi \left[\lambda_{\max}\{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}\} (\kappa\|\boldsymbol{\beta}\|_2 + \alpha)/\sigma \right] \right), \end{aligned}$$

where α is any positive constant and $\kappa = \lambda_{\min}(\boldsymbol{X}^T\boldsymbol{X})/\lambda_{\max}(\boldsymbol{X}^T\boldsymbol{X})$.

The equation (5.10) in Theorem 3 shows that the oracle estimator $\bar{\boldsymbol{\beta}}_{obs}^{(t)}$ is optimal and has lower Bayes risk than the OLS estimator $\hat{\boldsymbol{\beta}}_{obs}^{(t)}$ for large sample size n. Similarly as in Theorem 2, we can show that the proposed estimator $\widetilde{\boldsymbol{\beta}}_{obs}^{(t)}(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{W})$ converges in probability to the oracle estimator, $\left\| \widetilde{\boldsymbol{\beta}}_{obs}^{(t)}(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{W}) - \bar{\boldsymbol{\beta}}_{obs}^{(t)}(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{W}) \right\|_2 \xrightarrow{p} 0$, as $m \to \infty$. Thus, the proposed estimator is superior to the OLS estimator for missing data when the number of tissues is large enough.

6. Simulation

In this section, we conduct simulation studies to evaluate the performance of the proposed method and compare it with the OLS method, the unified test for molecular signatures (UTMOST) of Hu et al. (2019), the variational empirical Bayes (VEB) method of Kim et al. (2022), the nonparametric empirical Bayes (NEB) method of Wang and Zhao (2021), and the multiple regression with multivariate adaptive shrinkage (Mr.Mash) method of Morgante et al. (2023). The simulation results show that the proposed method achieves more accurate prediction and parameter estimation than the existing methods.

The R function of the proposed method has been made publicly available online at `https://github.com/feixue-stat/Multivariate-Empirical-Bayes`. The UTMOST method is implemented by codes in `https://github.com/ajmolstad/MTeQTLResults` (Molstad et al., 2021). We use the R packages "mr.ash.alpha" (`https://github.com/stephenslab/mr.ash.alpha`), "cole" (`https://github.com/sdzhao/cole`), and "mr.mash.alpha" (`https://github.com/stephenslab/mr.mash.alpha`) to implement the VEB, NEB, and Mr.Mash methods, respectively. Since the R function of Mr.Mash method encounters errors when the sample size is smaller than or equal to the number of tissues, we compare the proposed method with Mr.Mash method only under Settings 4 and 6 below.

In each simulation setting, 100 replications are performed. For each replication, we let

$$\boldsymbol{Y}^{(t)} = \boldsymbol{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{\varepsilon}^{(t)},$$

where $\boldsymbol{\varepsilon}^{(t)} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, and each row of the $n \times p$ matrix $\boldsymbol{X}$ is independent and identically distributed, for $t = 1, \ldots, m$. More details of simulation settings are provided in Settings 1-6 below. In each replication of each setting, we generate n training samples and n testing

samples.

To evaluate the performance of each method, we calculate the mean squared error (MSE) of each estimator $\widehat{\boldsymbol{\beta}}^{(t)}$ based on training sets and prediction MSE (PMSE) of the corresponding prediction $\widehat{\boldsymbol{Y}}^{(t)}(1 \leq t \leq m)$ in testing sets as follows:

$$\frac{1}{pm}\sum_{t=1}^{m}\|\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^{(t)}\|_2^2 \quad \text{and} \quad \frac{1}{nm}\sum_{t=1}^{m}\|\widehat{\boldsymbol{Y}}^{(t)}-\boldsymbol{Y}^{(t)}\|_2^2,$$

which measures the parameter estimation accuracy and prediction accuracy, respectively, across all tissues, where $\widehat{\boldsymbol{Y}}^{(t)}=\boldsymbol{X}_{test}\widehat{\boldsymbol{\beta}}^{(t)}$, and $\boldsymbol{X}_{test}$ represents the design matrix in testing data. We say that a method has better performance if the MSE and PMSE are smaller.

We compare the proposed method with existing methods under the following settings. In the first two settings, we assume that there is no missing values in the response $\boldsymbol{Y}^{(t)}$. The difference of the two settings mainly comes from the generation of $\boldsymbol{\beta}^{(t)}$. Specifically, we generate $\boldsymbol{\beta}^{(t)}$ from the proposed mixture prior distribution in Setting 1, while $\boldsymbol{\beta}^{(t)}$ is defined based on a low-rank matrix in Setting 2. For Settings 3 and 4, we consider the cases with missing values in the response, where we adopt the proposed estimator in Section 5 to deal with the missing responses. Moreover, in Setting 4, we directly use the cis-SNPs in the real GTEx genotype data as covariates $\boldsymbol{X}$ to capture the linkage disequilibrium structures of the genotype data. In Settings 5 and 6, we consider different number of tissues and different priors of the regression coefficients, respectively.

Setting 1. Let $p=30$, $n=50$, $m=50$, $\boldsymbol{\beta}=(\beta_s\mathbf{1}_{10}^T,(\beta_s/2)\mathbf{1}_{10}^T,\mathbf{0}_{10}^T)^T$, and $\sigma^2=1$, where β_s is the signal level. We generate each row of $\boldsymbol{X}$ from $N_p(\mathbf{0},\boldsymbol{C})$ and generate $\boldsymbol{\beta}^{(t)}$ independently

from

$$\boldsymbol{\beta}^{(t)} \mid I^{(t)} = 1 \sim N_p(\boldsymbol{\beta}, \boldsymbol{C}), \tag{6.11}$$
$$\boldsymbol{\beta}^{(t)} \mid I^{(t)} = 0 \equiv \mathbf{0},$$

for $t = 1, \ldots, m$, where $I^{(t)} \sim B(1, \tau_1)$ with $\tau_1 = 0.5$, and $\boldsymbol{C}$ is an exchangeable covariance matrix with all diagonals 1 and off-diagonals ρ.

Table 1: MSEs and PMSEs of different the methods under Setting 1. "MSE_OLS", "MSE_UTMOST", "MSE_VEB", "MSE_NEB" and "MSE_mEBmix" represent MSEs of the OLS, UTMOST, VEB, NEB, and the proposed method, respectively. "PMSE_OLS", "PMSE_UTMOST", "PMSE_VEB", "PMSE_NEB", "PMSE_mEBmix" represent PMSEs of the OLS, UTMOST, VEB, NEB, and the proposed method, respectively.

Correlation ρ	0			0.6			0.8		
Signal level β_s	0.5	1	2	0.5	1	2	0.5	1	2
MSE_OLS	0.053	0.053	0.052	0.129	0.127	0.128	0.253	0.254	0.256
MSE_UTMOST	0.273	0.369	0.770	0.265	0.314	0.612	0.286	0.354	0.611
MSE_VEB	0.028	0.028	0.033	0.076	0.074	0.087	0.388	0.417	0.498
MSE_NEB	0.557	0.704	1.255	0.333	0.382	0.444	0.234	0.277	0.368
MSE_mEBmix	0.025	0.025	0.025	0.067	0.062	0.063	0.128	0.130	0.130
PMSE_OLS	2.602	2.587	2.548	2.579	2.594	2.595	2.590	2.571	2.591
PMSE_UTMOST	9.267	11.989	24.091	14.298	16.889	35.220	13.733	16.135	25.143
PMSE_VEB	1.879	1.876	2.011	1.969	1.983	2.112	3.488	3.623	4.097
PMSE_NEB	17.804	22.185	38.969	48.151	45.157	29.260	39.451	39.228	31.108
PMSE_mEBmix	1.761	1.751	1.755	1.815	1.788	1.781	1.834	1.834	1.811

In Setting 1, we generate $\boldsymbol{\beta}^{(t)}$ following our model construction in Section 2.1, but we do not set the covariance matrix of $\boldsymbol{\beta}^{(t)}$ to be exactly $\eta(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Nevertheless, as shown in Table 1, the proposed method still performs better than existing methods under various correlation and signal levels. For example, when $\rho = 0.6$ and $\beta_s = 2$, the MSEs of the OLS,

UTMOST, VEB, NEB methods are 0.128, 0.612, 0.087, and 0.444, respectively, while the MSE of the proposed method is 0.063 which is only 49.2%, 10.3%, 72.4%, and 14.2% of that of the OLS, UTMOST, VEB, NEB methods, respectively. Moreover, the PMSEs of the OLS, UTMOST, VEB, NEB methods are 2.595, 35.220, 2.112, and 29.260, respectively, while the PMSE of the proposed method is only 1.781, which is 68.6%, 5.1%, 84.3, and 6.1% of that of the OLS, UTMOST, VEB, NEB methods, respectively.

Setting 2. We consider a similar setting as Setting 1 except that $\boldsymbol{\beta}^{(t)}$ is the t-th column of $\boldsymbol{AB}$ when $I^{(t)} = 1$ for $t = 1, \ldots, m$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are $p \times r$ and $r \times m$ matrices, respectively, whose elements are independently drawn from a standard normal distribution. We note that $\boldsymbol{\beta}^{(t)}$ is generated from the columns of $\boldsymbol{AB}$, which is different from our assumed mixture normal distribution in Section 2.1. Thus, the purpose of Setting 2 is to investigate robustness of the proposed method.

As shown in Table 2, the proposed method produces much smaller PMSE and MSE than the existing methods. For instance, at $\rho = 0.8$ and $r = 15$, the PMSE of the proposed method is only 69.5%, 5.3%, 5.3%, and 2% of that of the OLS, UTMOST, VEB and NEB, respectively. The MSE of the proposed method is at most 50% of those of existing methods. Thus, the proposed method is robust to certain errors in the model assumption.

In real genetic data such as the GTEx data, we might not collect all tissues from each subject, indicating that the gene expression matrix $\boldsymbol{Y}$ could contain missing values. We provided a proposed estimator in Section 5 to handle cases with missing values in $\boldsymbol{Y}$. In the following Settings 3 and 4, we consider missing responses, where we compare the proposed method with only the OLS and the UTMOST methods since the VEB and the NEB methods are unable to deal with missing data.

Table 2: MSEs and PMSEs of all the methods under Setting 2. "MSE_OLS", "MSE_UTMOST", "MSE_VEB", "MSE_NEB", "MSE_mEBmix" represent MSEs of the OLS, UTMOST, VEB, NEB, and the proposed method, respectively. "PMSE_OLS", "PMSE_UTMOST", "PMSE_VEB", "PMSE_NEB", "PMSE_mEBmix" represent PMSEs of the OLS, UTMOST, VEB, NEB, and the proposed method, respectively.

Correlation ρ	0			0.6			0.8		
Rank r	5	10	15	5	10	15	5	10	15
MSE_OLS	0.052	0.053	0.053	0.128	0.126	0.126	0.263	0.256	0.252
MSE_UTMOST	1.265	2.541	3.751	1.549	2.910	4.607	1.649	3.259	4.988
MSE_VEB	0.042	0.062	0.081	0.432	0.700	1.110	1.705	3.307	5.162
MSE_NEB	2.406	4.829	6.272	2.463	4.278	6.209	2.407	4.424	6.444
MSE_mEBmix	0.026	0.027	0.026	0.065	0.060	0.063	0.128	0.124	0.126
PMSE_OLS	2.552	2.618	2.587	2.603	2.576	2.577	2.620	2.569	2.561
PMSE_UTMOST	39.016	77.867	115.268	21.277	38.524	60.730	12.166	22.103	33.590
PMSE_VEB	2.284	2.932	3.533	6.527	9.800	15.032	11.677	21.605	33.384
PMSE_NEB	73.767	147.736	192.539	68.910	102.722	128.812	59.925	76.017	87.425
PMSE_mEBmix	1.766	1.809	1.773	1.813	1.760	1.788	1.795	1.766	1.781

Setting 3. We follow similarly as in Setting 1 except $\tau_1 = 0.1$ and that, we randomly set 20% elements in $\boldsymbol{Y}^{(t)}$ to be missing for each $t = 1, \ldots, m$.

We provide results of Setting 3 in Table D.1 in Supplementary Materials. Under Setting 1, we let $\tau_1 = 0.5$, indicating that the two groups of tissues ($\boldsymbol{\beta}^{(t)} \neq \mathbf{0}$ verse $\boldsymbol{\beta}^{(t)} = \mathbf{0}$) are balanced, while in Setting 3, we provide simulations with $\tau_1 = 0.1$ where the two groups of tissues are unbalanced. Due to the imbalance and missing values, the PMSEs and MSEs of the OLS in Setting 3 are larger than those in Setting 1. In contrast, the PMSEs and MSEs of UTMOST and the proposed method in Setting 3 are smaller than those in Setting 1, since these two methods are designed for the cases with sparse coefficients. As shown in Table D.1, the proposed method outperforms the OLS and UTMOST methods in terms of

both PMSE and MSE.

Setting 4. We follow similarly as in Setting 3 except $n = 300$ and that, for the construction of $\boldsymbol{X}$, we randomly select n samples independently with replacement from the 838 samples of the 30 cis-SNPs of the gene *WARS2* in the GTEx data.

The results under Setting 4 are provided in Table D.2 in Supplementary Materials, showing that the proposed method still produces smaller PMSEs and MSEs than existing methods when we use the real cis-SNP genotype covariates in GTEx. In particular, the proposed method performs better than Mr.Mash under the settings with various noise levels, correlations, and signal levels. For example, with $\sigma = 2$, $\rho = 0.8$ and $\beta_s = 0.5$, the PMSE of the proposed estimator is 4.094, which is smaller than PMSEs of the OLS, UTMOST, and Mr.Mash methods. Note that, as the standard deviation (σ) of the error term increases, signal level (β_s) increases, or correlation (ρ) increases, the difference between PMSEs of UTMOST and the proposed method becomes larger in most cases, while the PMSEs of the proposed method across correlations and signal levels do not change much. This indicates that the proposed method is more stable in terms of PMSE compared to the UTMOST method.

In addition to MSE and PMSE, in Settings 1-4, we use the posterior probability of $I^{(t)}$ produced by the proposed method to calculate the area under the receiver operating characteristic curve (AUC) for identifying the tissues with non-zero cis-effects. The AUCs of the proposed method across different settings are all above 0.95 and most are very close to 1, indicating that the proposed method can effectively find tissues with non-zero $\boldsymbol{\beta}^{(t)}$.

Since the performance of the proposed method depends on the number of tissues m based on Theorem 2 in Section 4, we choose different m values in the following setting to

illustrate this point.

Setting 5. We follow a similar setting as Setting 1 except that $\beta_s = 0.5$, $\sigma^2 = 5$, $m = 3, 30,$ or 300, and that the covariance matrix in equation (6.11) is $\boldsymbol{C}/10$.

The results of the proposed method are provided in Table 3. The MSE and PMSE both decrease as the number of tissues increases under different correlation settings, indicating that the proposed method performs better when we have more tissues.

Table 3: MSE and PMSE of the proposed method mEBmix under Setting 5.

Correlation ρ	0			0.4			0.6		
Number of tissues m	3	30	300	3	30	300	3	30	300
MSE	0.525	0.116	0.063	0.833	0.145	0.100	1.227	0.272	0.196
PMSE	40.812	28.377	26.865	42.522	29.658	28.768	42.239	31.703	30.848

In the following Setting 6, we compare the methods under different priors of the regression coefficients.

Setting 6. We follow similarly as in Setting 1 except for $n = 100$, $m = 10$, $\beta_s = 0.5$, $\rho = 0.2$, $\sigma^2 = 2$, and the prior distribution of coefficients $\boldsymbol{\beta}^{(t)}$ for $t = 1, \ldots, m$. We consider the following priors in this setting.

- Proposed prior in equations (2.3) and (2.4) with $\eta = 0.1$ and $\boldsymbol{\beta} = (0.5 \times \mathbf{1}_{10}^T, 0.25 \times \mathbf{1}_{10}^T, \mathbf{0}_{10}^T)^T$.
- Multivariate normal mixture (MNM) prior: $\boldsymbol{\beta}_j \overset{i.i.d.}{\sim} 0.5N(\mathbf{0}, \boldsymbol{I}_m) + 0.5N(\mathbf{0}, \mathbf{1}_m\mathbf{1}_m^T)$ for $j = 1, \ldots, p$, where $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_j^{(1)}, \ldots, \boldsymbol{\beta}_j^{(m)})^T$ is a m-dimensional coefficient vector corresponding to the j-th covariate and $\boldsymbol{\beta}_j^{(t)}$ is the j-th element in $\boldsymbol{\beta}^{(t)}$ for $t = 1, \ldots, m$.
- Univariate normal mixture (UNM) prior: $\boldsymbol{\beta}_j^{(t)} \overset{i.i.d.}{\sim} 0.5N(0, \sigma_1^2) + 0.5N(0, \sigma_2^2)$ with $\sigma_1^2 = 1$ and $\sigma_2^2 = 0$ for $j = 1, \ldots, p$ and $t = 1, \ldots, m$.

In this setting, we consider not only the prior that is used in the proposed method, but also the multivariate normal mixture (MNM) prior and the univariate normal mixture (UNM) prior that are assumed in Morgante et al. (2023) and Kim et al. (2022) for Mr.Mash and VEB methods, respectively. We do not consider any particular priors for the OLS, UTMOST, and NEB methods, since these methods do not assume any specific prior of coefficients.

The results of Setting 6 are provided in Table 4, which show that the proposed method performs the best among all the methods in terms of both MSE and PMSE when the proposed prior is used to generate the regression coefficients. When the multivariate normal mixture prior and the univariate normal mixture prior are used, Mr.Mash and VEB methods perform the best, respectively. Thus, each method achieves the best when data are generated from its own model. Nevertheless, the proposed method still performs well and is the second best in terms of MSE and PMSE, when coefficients are not generated from the proposed prior, indicating certain robustness of mEBmix against the prior.

We provide computation time of each method for one replication under Setting 6 in Table 4. It shows that, the proposed method takes longer than the OLS, VEB, and NEB methods, but is faster than the UTMOST and Mr.Mash methods. This indicates that the computation time of the proposed method is comparable to that of existing methods.

7. Application to GTEx data

In this section, we apply the proposed method to the Genotype-Tissue Expression (GTEx) data (Consortium et al., 2017) and compare it with the OLS, the UTMOST, and Mr.Mash methods in terms of predicting tissue-specific expressions using the cis-SNP data. In Section

Table 4: MSEs, PMSEs, and computation time of OLS, UTMOST, VEB, NEB, Mr.Mash, and the proposed method mEBmix under Setting 6. "Proposed", "MNM", and "UNM" represent the proposed prior, multivariate normal mixture prior, and univariate normal mixture prior, respectively. For each method. "MSE_" represents MSE, "PMSE_" represents PMSE, and "Time_" represents computation time (in seconds) for one replication.

Prior	Proposed	MNM	UNM
MSE_OLS	0.070	0.034	0.034
MSE_UTMOST	0.023	0.404	0.088
MSE_VEB	0.018	0.034	0.022
MSE_NEB	0.082	0.680	0.693
MSE_Mash	0.036	0.017	0.055
MSE_mEBmix	0.008	0.032	0.033
PMSE_OLS	5.769	2.848	2.859
PMSE_UTMOST	5.619	12.158	4.186
PMSE_VEB	4.525	2.885	2.581
PMSE_NEB	9.337	21.495	21.580
PMSE_Mash	6.299	2.482	3.387
PMSE_mEBmix	4.207	2.813	2.821
Time_OLS	0.001	0.001	0.001
Time_UTMOST	3.230	3.138	3.212
Time_VEB	0.306	0.350	0.354
Time_NEB	0.002	0.001	0.000
Time_Mash	5.553	1.033	3.412
Time_mEBmix	0.996	0.584	0.521

6, we also compared the proposed method with the VEB and the NEB methods. However, they are unable to deal with missing data, and thus they are not used in the real data application.

We are interested in understanding the cis-SNP and gene expression associations across different issues. The GTEx is an ongoing US National Institutes of Health (NIH) Common Fund project starting from 2010, which aims to establish a comprehensive public resource database for investigation of the relationship between genetic variation and gene expression. The GTEx project collects non-diseased tissue samples from nearly 1000 donors, which are sent to the Laboratory, Data Analysis and Coordinating Center (LDACC) for molecular analysis (Lonsdale et al., 2013). DNA from each donor's blood sample is genotyped using the Illumina HumanOmni5M-Quad BeadChip for whole-genome SNP (Lonsdale et al., 2013), and the Illumina TrueSeq RNA sequencing is used for the measurement of gene expression.

Specifically, the GTEx project includes genotype and gene expression data of 838 participants across 49 tissue types. We extract the gene expression values from "GTEx_Analysis_v8_eQTL_expression_matrices.tar" under "QTL" tab at `https://www.gtexportal.org/home/downloads/adult-gtex`, which are fully processed, filtered, and normalized by the GTEx project (Lonsdale et al., 2013; GTEx Consortium, 2020). Details about the pre-processing of the expression data can be found in Section 3 of the Supplementary Material of GTEx Consortium (2020). There are 8066 genes with expression data available in all the tissue types. Since tissue samples of some participants are not collected for each tissue type, our analyses of GTEx data focus on 32 tissues, each of which has at least $n = 200$ collected samples.

For DNA data, we use the GTEx genotype dataset "GTEx_Analysis_2017-06-05_v8_WholeGenomeSeq_838Indiv_Analysis_Freeze.SHAPEIT2_phased.vcf.g", whose quality-control pro-

cedure was also conducted by the GTEx study. More details can be found in Section 2 of the Supplementary Material of GTEx Consortium (2020). The genotype data are GTEx protected access data, which can be obtained by following the steps at `https://www.gtexportal.org/home/protectedDataAccess`.

We further process the genotype dataset as follows. We first exclude the SNPs with minor allele frequencies less than 5%. The SNPs are further pruned for LD with a window size of 50 SNPs, a step size of 5 SNPs, and a R2 threshold of 0.2 using PLINK 1.9. In addition, we select cis-eQTLs for each gene following Wang et al. (2016), which can be viewed as a screening of predictors. Specifically, we first obtain tissue-specific cis-eQTLs of pairs of genes and its corresponding cis-SNPs using the "MatrixEQTL" R package (`https://cran.r-project.org/web/packages/MatrixEQTL/index.html`). For each pair, we then combine the Z statistics of the cis-eQTLs from all the tissues via the Stouffer's Method (Stouffer et al., 1949). We use the "poolr" R package (`https://cran.r-project.org/web/packages/poolr/index.html`) to carry out the Stouffer's Method. We select the cis-eQTLs whose Stouffer's p values are less than 10^{-6}. Then there are 4827 genes with at least one selected cis-eQTL. For each gene, we order the selected cis-SNPs by the corresponding Stouffer's p values increasingly. To avoid highly correlated cis-SNPs, we retain the selected cis-SNPs in a increasing order of the corresponding Stouffer's p values, and remove cis-SNPs which are highly correlated with previously retained cis-SNPs with the coefficient of determination larger than 0.5.

We apply the proposed method, OLS, UTMOST method, and Mr.Mash method to each gene and its corresponding cis-SNPs. To evaluate each method in gene expression prediction based on cis-SNPs, we use a 10-fold cross-validation analysis. Specifically, we

randomly split the observed samples in each tissue into 10 equally sized subsamples, named from Subsample 1 to Subsample 10. For each $1 \leq i \leq 10$, we use Subsample i in all the tissues as a testing set, and the remaining 9 subsamples in all the tissues as a training set. We predict the gene expression values of subjects in each testing set for each gene based on each method. For the proposed method, we use the extension version in Section 5 that can handle missing values, since there are missing samples in some tissues for a subject in the training set. We repeat this procedure 10 times and obtain predicted values for all the subjects. To evaluate the prediction accuracy of each method, we calculate the prediction mean squared error and Pearson correlation between the predicted values and true gene expression values for each gene and each tissue type.

For each tissue type, we take averages of PMSEs and correlations, respectively, across all the genes. The results are provided in Tables E.1 and E.2 in Supplementary Materials, showing that the proposed method produces the smallest PMSE and the highest correlation in each tissue type among all the methods. Among all the tissue types, the improvement of PMSE by the proposed mEBmix is relatively higher in "Adrenal_Gland", "Brain_Nucleus_accumbens_basal_gan glia", "Colon_Transverse", "Pancreas", "Pituitary", and "Spleen" tissues.

For these tissue types, we also provide the PMSEs of all genes by each method in Figures 1 and 2. Each sub-figure in Figures 1 and 2 is a scatter plot of PMSEs of all genes in one tissue type, where x-axis represents PMSE of an existing method, and y-axis represents PMSE of the proposed method. In each sub-figure, the majority of points are under the red line where the PMSEs of the two methods are the same. Thus, the proposed method overall performs better than existing methods in terms of PMSE. Especially, in most of

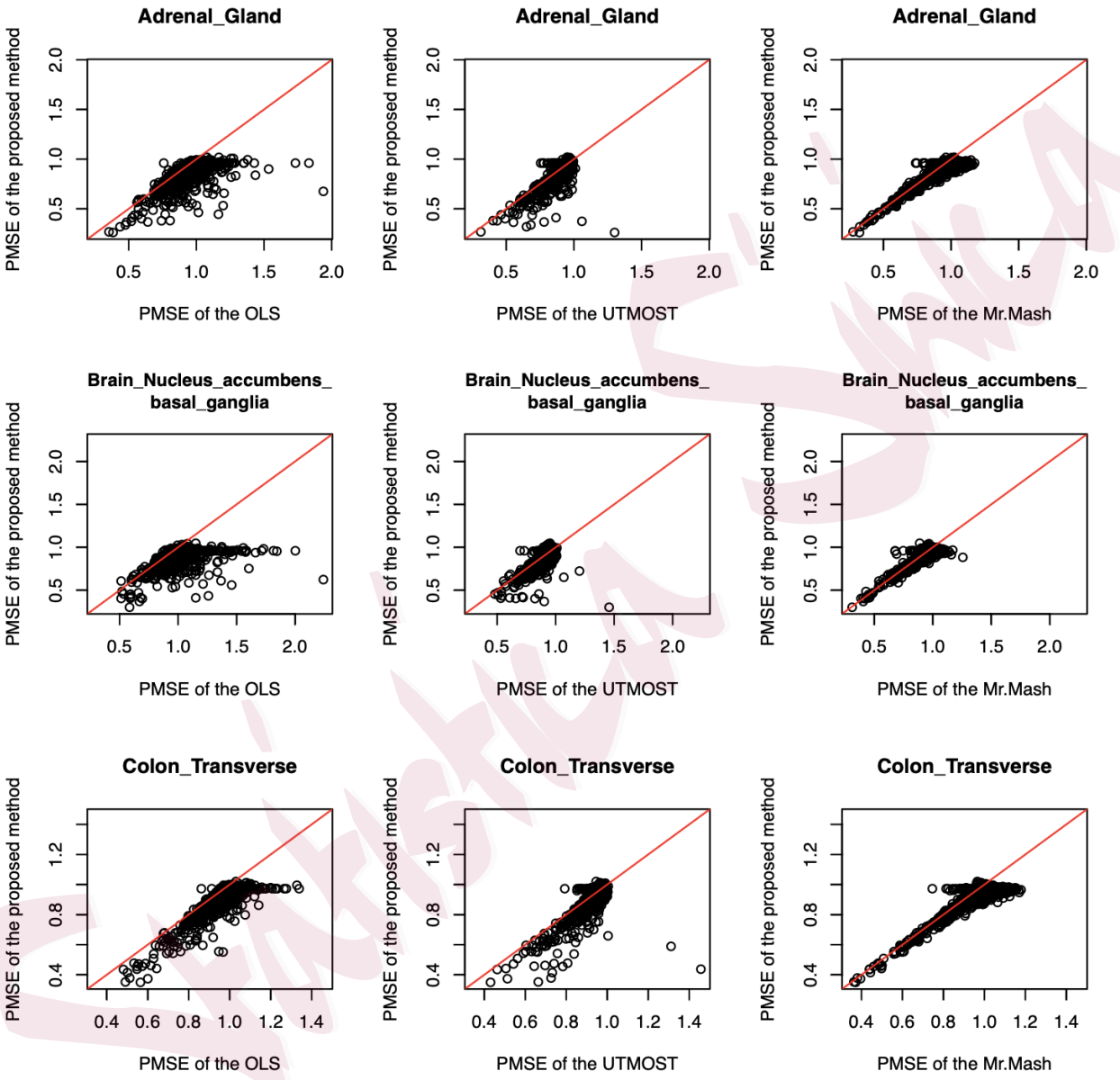

Figure 1: Scatter plots of PMSEs in various tissues. Y-axis: PMSE of the proposed method. X-axis: PMSE of OLS, UTMOST, and Mr.Mash, respectively.

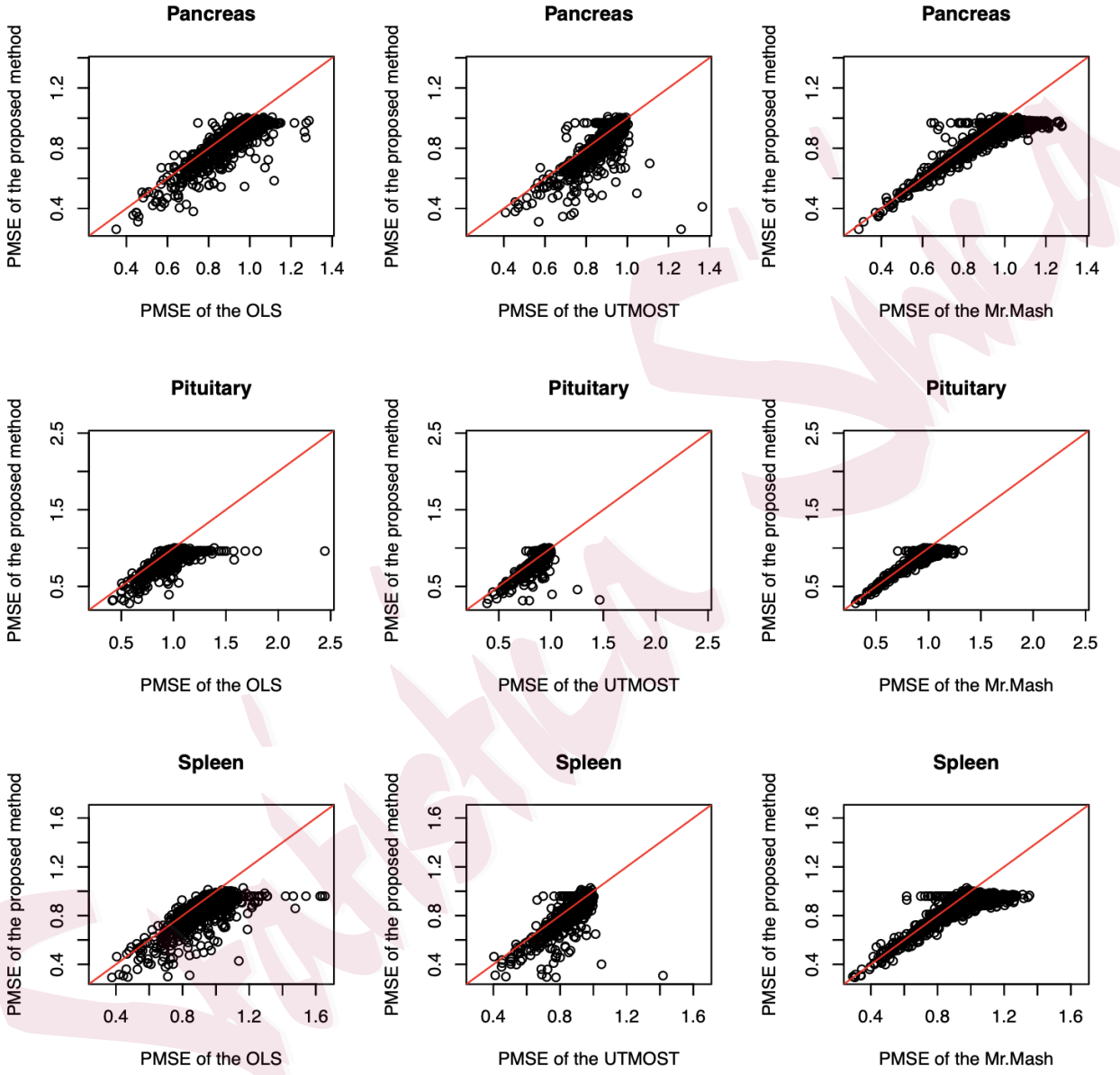

Figure 2: Scatter plots of PMSEs in various tissues. Y-axis: PMSE of the proposed method. X-axis: PMSE of OLS, UTMOST, and Mr.Mash, respectively.

these tissues, for some genes, the PMSEs of the proposed method is even smaller than 50% of the corresponding PMSEs from UTMOST.

Moreover, in Figure 3, we plot the absolute increase and relative increase of correlation by the proposed method compared with Mr.Mash method, where the absolute increase is the difference between average correlations of the proposed method and Mr.Mash method, and the relative increase is the absolute increase divided by the average correlation of Mr.Mash method. We observe that the proposed method mEBmix increases the average correlations across genes in all the tissues compared to Mr.Mash method, indicating that the predicted values from mEBmix are more correlated to the true gene expression values. On average, the proposed method improves the correlation by 34.9% across all the tissue types. In particular, mEBmix increases the correlation by over 100% in the "Brain_Nucleus_accumbens_basal_ganglia" tissue.

Furthermore, we apply the proposed method to the whole dataset, calculate the posterior probability of $I^{(t)} = 1$ for all tissues and genes, and generate a heat map of all the posterior probabilities in Section E.4 of the Supplementary Materials. The heat map indicates that the posterior probabilities based on the proposed method indeed capture the similarity between tissues in terms of the relationship between gene expression and cis-SNPs.

Since mEBmix assumes that gene expression values follow a normal distribution, we use Shapiro–Wilk test to check the normality of expression values of each gene and each tissue that we used in our real data application. Among the 4827 genes, there are only 368 genes whose corresponding p-values are smaller than 0.05 in at least one tissue. Moreover, in these 368 genes, there are 251 genes whose p-values are smaller than 0.05 in only one tissue. Thus, the expression values of most genes can be regarded as being normally distributed in

most tissues.

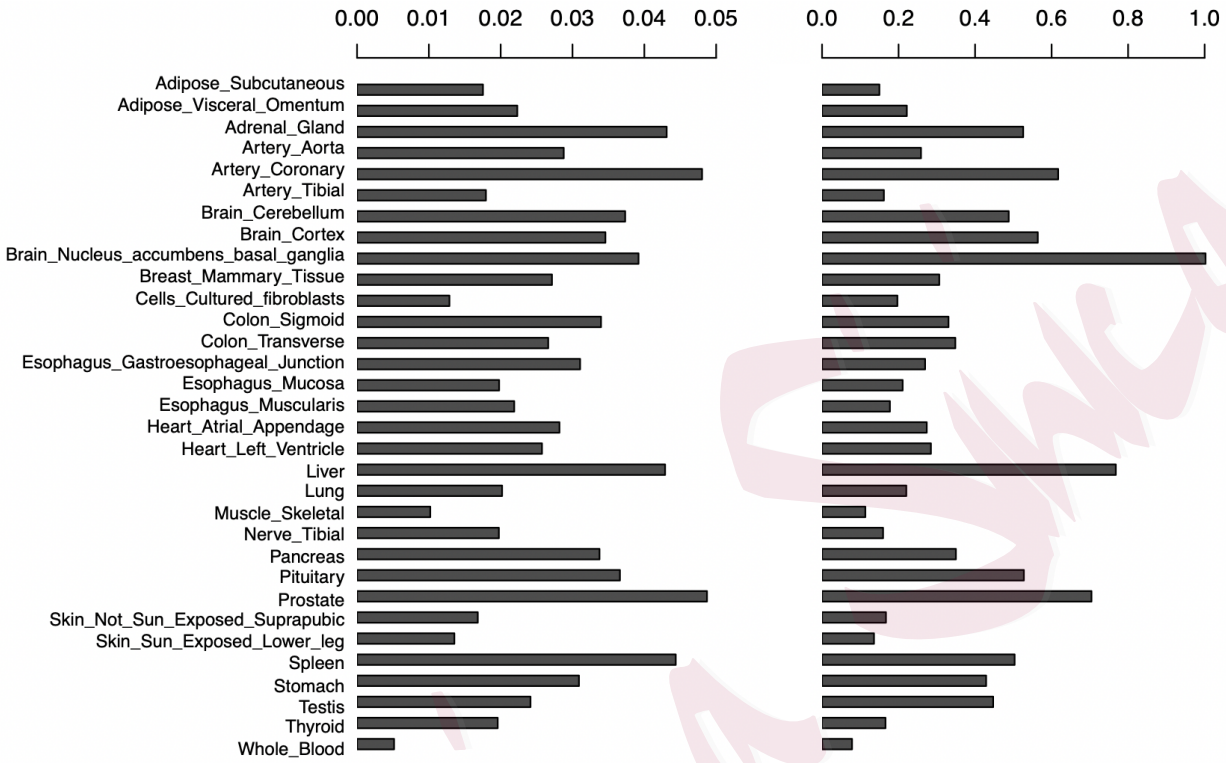

Figure 3: Increase of correlation by the proposed method compared with Mr.Mash method. Left: Increase of average correlation by the proposed method. Right: Relative increase of correlation by the proposed method.

To check robustness of the proposed method against the normal distribution assumption, for each method and each tissue, we take average of prediction mean squared errors (PMSEs) of genes whose p values are smaller than 0.05 in at least 20 tissues. That is, expression values of these genes are not normally distributed in most of tissues according to the Shapiro–Wilk test. The average PMSEs are provided in Table E.3 in Supplementary Materials. We observe that the PMSE of the proposed method is smaller than that of other methods in each tissue. This indicates that the proposed method still performs the best among all the methods even when the responses, gene expression values, are not normally distributed.

We finally provide the computation time (in seconds) of all the methods for genes with different number of cis-SNPs in Table E.4 in Supplementary Materials. Specifically, we record the running time of the OLS, UTMOST, Mr.Mash, and the proposed method for a gene with the number of cis-SNPs (or predictors) $p = 30, 50, 80, 105,$ or 125. For each p, this gene is randomly selected from all the genes with p cis-SNPs (or predictors). The results show that the computation time increases as the number of predictors increases in most cases. The proposed method takes longer than the OLS method, but mostly shorter than UTMOST and Mr.Mash, which is consistent with the results on computation time under simulation Setting 6 in Section 6.

8. Discussion

We develop a new empirical Bayes regression model for SNP-based gene-expression prediction and multi-tissue eQTL analysis. To borrow information across tissues, the proposed method assigns a common mixture prior distribution to the cis-SNP effects in each tissue, and estimates parameters in the prior distribution through maximizing marginal likelihood of the expression levels in all the tissues. In addition, the method provides a way of quantify the evidence whether the cis-SNPs are "active" or not in a certain tissue based on the posterior probabilities of the latent configuration indicator in the mixture prior distribution. We apply the EM algorithm to find the maximum likelihood estimate of prior parameters. Moreover, to accommodate real data with missing responses such as the GTEx data, we have also developed the empirical Bayes estimator and the corresponding asymptotic results for missing data.

Theoretically, we have shown that the proposed estimator is asymptotically superior

than the OLS estimator in terms of the Bayes risk. This superiority is mainly due to that the OLS only uses single tissue information while the proposed method incorporates common information from other tissues. Moreover, we have demonstrated that the proposed method converges to the optimal oracle estimator as the number of tissue types increases. In addition, the application to the GTEx data illustrates that the proposed method predicts gene expression more accurately than existing methods. More importantly, the proposed method provides posterior probabilities of whether there is cis-effects or not for each tissue, which indeed reflects similarity among tissues.

In general, the empirical Bayes method provides a powerful framework for pooling information across multiple experiments or sources, and improving the accuracy of the estimation or inference in each experiment. Besides the SNP-gene association, we can also extend the empirical Bayes framework to improve the estimation of the relationship between expression levels of genes for the GTEx project where gene expression levels are measured over multiple tissues. For example, we could incorporate information across tissues through estimating a common prior on the multiple precision matrices for the multiple tissues. In addition, in this article, we mainly consider the association between gene expression and cis-SNPs. It would be of great interest to incorporate more covariates, including not only cis-SNPs but also trans-SNPs, in future research. We could involve penalty functions when the number of covariates exceeds the number of subjects.

Acknowledgments

The authors thank Jianqiao Wang for his help on processing the GTEx data. This research was supported by NIH grants GM123056 and GM129781, and NSF grant DMS-2210860.

References

Boyle, E. A., Y. I. Li, and J. K. Pritchard (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell 169*(7), 1177–1186.

Brem, R. B., J. D. Storey, J. Whittle, and L. Kruglyak (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature 436*(7051), 701–703.

Consortium, G. et al. (2017). Genetic effects on gene expression across human tissues. *Nature 550*(7675), 204–213.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B 39*(1), 1–22.

Dezső, Z., Y. Nikolsky, E. Sviridov, W. Shi, T. Serebriyskaya, D. Dosymbekov, A. Bugrim, E. Rakhmatulin, R. J. Brennan, A. Guryanov, et al. (2008). A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biology 6*(1), 1–15.

Duong, D., L. Gai, S. Snir, E. Y. Kang, B. Han, J. H. Sul, and E. Eskin (2017). Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of egenes. *Bioinformatics 33*(14), i67–i74.

Duong, D., J. Zou, F. Hormozdiari, J. H. Sul, J. Ernst, B. Han, and E. Eskin (2016). Using genomic annotations increases statistical power to detect egenes. *Bioinformatics 32*(12), i156–i163.

Flutre, T., X. Wen, J. Pritchard, and M. Stephens (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics 9*(5), e1003486.

Gamazon, E. R., H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, D. L. Nicolae, N. J. Cox, et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics 47*(9), 1091.

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Sci-*

ence 369(6509), 1318–1330.

Gusev, A., A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx, R. Jansen, E. J. De Geus, D. I. Boomsma, F. A. Wright, et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics 48*(3), 245–252.

Hu, Y., M. Li, Q. Lu, H. Weng, J. Wang, S. M. Zekavat, Z. Yu, B. Li, J. Gu, S. Muchnik, et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics 51*(3), 568–576.

Kim, Y., W. Wang, P. Carbonetto, and M. Stephens (2022). A flexible empirical Bayes approach to multiple linear regression and connections with penalized regression. *arXiv preprint arXiv:2208.10910*.

Li, G., D. Jima, F. A. Wright, and A. B. Nobel (2018). HT-eQTL: Integrative expression quantitative trait loci analysis in a large number of human tissues. *BMC Bioinformatics 19*, 1–11.

Lonsdale, J., J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, et al. (2013). The genotype-tissue expression (GTEx) project. *Nature Genetics 45*(6), 580–585.

Molstad, A. J., W. Sun, and L. Hsu (2021). A covariance-enhanced approach to multi-tissue joint eQTL mapping with application to transcriptome-wide association studies. *The Annals of Applied Statistics 15*(2), 998.

Morgante, F., P. Carbonetto, G. Wang, Y. Zou, A. Sarkar, and M. Stephens (2023). A flexible empirical Bayes approach to multivariate multiple regression, and its improved accuracy in predicting multi-tissue gene expression from genotypes. *PLoS Genetics 19*(7), e1010539.

Shi, X., X. Chai, Y. Yang, Q. Cheng, Y. Jiao, H. Chen, J. Huang, C. Yang, and J. Liu (2020). A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. *Nucleic Acids Research 48*(19), e109–e109.

Sonawane, A. R., J. Platig, M. Fagny, C.-Y. Chen, J. N. Paulson, C. M. Lopes-Ramos, D. L. DeMeo, J. Quackenbush, K. Glass, and M. L. Kuijjer (2017). Understanding tissue-specific gene regulation. *Cell Reports 21*(4), 1077–1088.

Stegle, O., L. Parts, M. Piipari, J. Winn, and R. Durbin (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols 7*(3), 500.

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr (1949). *The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1.* Princeton Univ. Press.

Stranger, B. E., A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, et al. (2007). Population genomics of human gene expression. *Nature Genetics 39*(10), 1217–1224.

Sul, J. H., B. Han, C. Ye, T. Choi, and E. Eskin (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genetics 9*(6), e1003491.

Uffelmann, E., Q. Q. Huang, N. S. Munung, J. De Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma (2021). Genome-wide association studies. *Nature Reviews Methods Primers 1*(1), 1–21.

Wainberg, M., N. Sinnott-Armstrong, N. Mancuso, A. N. Barbeira, D. A. Knowles, D. Golan, R. Ermel, A. Ruusalepp, T. Quertermous, K. Hao, et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics 51*(4), 592–599.

Wang, J., E. R. Gamazon, B. L. Pierce, B. E. Stranger, H. K. Im, R. D. Gibbons, N. J. Cox, D. L. Nicolae, and L. S. Chen (2016). Imputing gene expression in uncollected tissues within and beyond GTEx. *The American Journal of Human Genetics 98*(4), 697–708.

Wang, Y. and S. D. Zhao (2021). A nonparametric empirical Bayes approach to large-scale multivariate regression. *Computational Statistics & Data Analysis 156*, 107130.

Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

E-mail: feixue@purdue.edu

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA

E-mail: hongzhe@pennmedicine.upenn.edu